

(19)



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11)

EP 0 674 263 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention
of the grant of the patent:
25.11.1998 Bulletin 1998/48

(51) Int Cl.⁶: **G06F 11/20, G06F 11/14**

(21) Application number: **95301345.5**

(22) Date of filing: **02.03.1995**

(54) Asynchronous remote data copying

Asynchrones Datenfernkopieren

Replication de données distante et asynchrone

(84) Designated Contracting States:
DE FR GB

(30) Priority: **21.03.1994 US 215238**

(43) Date of publication of application:
27.09.1995 Bulletin 1995/39

(73) Proprietor: **International Business Machines
Corporation**
Armonk, N.Y. 10504 (US)

(72) Inventors:
• **Micka, William Frank**
Tucson, Arizona 85718 (US)
• **Shomler, Robert Wesley**
Morgan Hill, California 95037 (US)

(74) Representative: **Davies, Simon Robert**
I B M
UK Intellectual Property Department
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(56) References cited:
EP-A- 0 455 922 **WO-A-91/20034**

- **PROCEEDINGS OF THE ACM SIGMOD
INTERNATIONAL CONFERENCE ON
MANAGEMENT OF DATA, SAN DIEGO, JUNE 2 -
5, 1992, vol. 21, STONEBRAKER M, pages
236-245, XP 000393592 ANUPAM BHIDE ET AL
'AN EFFICIENT SCHEME FOR PROVIDING HIGH
AVAILABILITY'**

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 0 674 263 B1

Description

The present invention relates to asynchronous remote data copying or duplexing to provide data preservation in an information handling system, whereby information from a primary site storage subsystem can be copied to a remote location.

Data copying is one form of data preservation in an information handling or computer system. However, data preservation via data copying must take many factors into account. This is of special significance where it is anticipated that data copied and stored at a remote site would be the repository for any continued interaction with the data should the work and data of a primary site become unavailable. The factors of interest in copying include the protection domain (system and/or environmental failure or device and/or media failure), data loss (no loss/partial loss), time where copying occurs as related to the occurrence of other data and processes (point in time/real time), the degree of disruption to applications executing on said computer, and whether the copy is application or storage system based. With regard to the last factor, application based copying involves log files, data files, and program routines while storage based copying involves an understanding of direct access storage device (DASD) addresses with no knowledge of data types or application use of the data.

Real-time remote data duplexing systems require some means to ensure update sequence integrity as write updates to the secondary or remote DASD data copy. One way to accomplish this is to provide a synchronous system to control the DASD subsystems. In such a system, the primary DASD write operation does not complete until a copy of that data has been confirmed at a secondary location. The problem with such synchronous systems is that they slow down the overall operation of the duplexing system.

Asynchronous copy systems accomplish sequence integrity through a centralization and consolidation of data communications between primary and secondary DASD subsystems through a central communications system. In such systems, a system at the primary site can determine the sequence among different update write operations among all DASD subsystems at the primary site and communicate that information to the DASD subsystem at the remote site. The secondary subsystem in turn uses the sequence information from the primary to control the application of update data to the secondary DASD data copy. Known asynchronous copy systems that utilize centralized data communications are described below.

McIlvain and Shorler, U.S. patent application number 08/036,017 entitled "Method and Means for Multi-System Remote Data Duplexing and Recovery" (EPA 617362) describes the use of a store and forward message interface at the DASD storage management level between a source of update copies and a remote site in a host to host coupling in which the difference in

update completeness or loss of the sequence of write updates could be completely specified in the event of interruption.

Cheffetz, et al., U.S. Patent 5,133,065 entitled "Backup Computer Program for Networks" issued July 21, 1992, discloses a local area network (LAN) having a file server to which each local node creates and transmits a list of local files to be backed-up. Such remote generation reduces the traffic where a network server initiates the list creation and file copying activity. Arguably, art published before this reference taught centrally administered file selection. This resulted in compromises to local node security and overuse of the server. This is presumptively avoided by Cheffetz's local node generated lists and remission of the lists to the file server.

Beale, et al., U.S. Patent 5,155,845 entitled "Data Storage System for Providing Redundant Copies of Data on Different Disk Drives", copies variable length records (CKD) on two or more external stores by causing a write to be processed by the first storage controller and be communicated in parallel over a direct link (broad band path) to the second storage controller obviating the path length limitation between the primary and remote copy sites. Such a limitation is occasioned by the fact that CKD demand/response architecture is length limited to in the range of 150 meters.

Another example of an asynchronous system that utilizes a centralized system is disclosed in U.S. patent application entitled "Remote Data Duplexing Asynchronous Information Packet Message, by Micka et. al. (EPA 602822). This system discloses a system for asynchronously duplexing direct access storage device (DASD) data in a plurality of DASD subsystems and has the advantage of decoupling the data duplexing operation from the DASD write I/O operation. This ensures the write does not incur unnecessary wait states in the subsystem. By establishing a sequence checkpoint at which time a set of information packets are grouped together and processed as a single sequence unit, this decoupling and independent operation takes place. Through this independence, data copying to a secondary location can take place without affecting the performance of the subsystems and also not affecting the corresponding integrity of the data that is being updated.

EP-A-602822 describes a system for asynchronously duplexing direct access storage device (DASD) data in a plurality of DASD subsystems. Advantageously, the system decouples the data duplexing operation from the DASD write I/O operation, thereby ensuring that a write does not incur unnecessary wait states in a subsystem. The system uses a sequence checkpoint as part of this decoupling process.

There are systems in use in which there is no centralized communication service between the primary and secondary locations. Oftentimes in such system configurations each primary subsystem has a direct independent link to a selected secondary subsystem. In such a system that includes subsystems providing se-

quence consistent asynchronous write operations can not be addressed utilizing known asynchronous copy schemes.

Accordingly, the invention provides a method of operating an asynchronous remote data copy system including a primary site having a first plurality of storage subsystems interconnected by a first coupling means, and a secondary site remote from the primary site having a second plurality of storage subsystems interconnected by a second coupling means, each of the second plurality of storage subsystems being independently coupled to a counterpart one of the first plurality of storage subsystems, said method comprising the steps of:

in the first plurality of subsystems;
 sending a checkpoint signal to each of the first plurality of subsystems;
 sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and
 in the second plurality of subsystems;
 receiving and buffering the data updates and checkpoint signals;
 and coordinating the writing of the data updates based upon the checkpoint signals such that update sequence integrity is preserved at the secondary site.

Such method provides independent links between primary and secondary subsystems, with no central communications system, and provides for sequence consistent remote copying from one set of multiple subsystems at one location to a second set of multiple subsystems at a second subsystem. The asynchronous copy system is simple, cost effective and does not significantly impede the overall operation of the subsystems.

In a preferred embodiment, the step of sending a checkpoint signal to each of the first plurality of subsystems further comprises the step of distributing a sequence signal at the primary site, said checkpoint signal including information indicative of the value of the sequence signal. Such a sequence signal is used to ensure synchronisation between the various subsystems. However, if the network can ensure quick and reliable distribution of the checkpoint messages, then in some situations it is possible to dispense with the sequence signal.

It is further preferred that said step of sending data updates and a checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises the steps of: generating asynchronously a sequence of data updates at each of said first plurality of subsystems; ordering each of said sequences in accordance with the distributed sequence signal and checkpoint signal; and asynchronously communicating sequences of updates from each of the first plurality of subsystems to a counterpart subsystem.

In a preferred embodiment, the method further comprises the initial steps at the primary site of: activating each of the first plurality of subsystems to communicate with the other subsystems in said first plurality of subsystems; activating each of the first plurality of subsystems to communicate with its counterpart coupled subsystem in the second plurality of subsystems; building at least one configuration table in each of the first plurality of subsystems such that each of the first plurality of subsystems can identify all of the other subsystems in said first plurality of subsystems; and synchronizing the first plurality of subsystems;

and the method further comprises the initial steps at the secondary site of: receiving copy active messages in the second plurality of subsystems; building copy active tables in the second plurality of subsystems; and synchronizing copy operations of the second plurality of subsystems;
 and wherein the coordinating step further comprises the step of performing a rendezvous for all checkpoint signals in the second plurality of subsystems.

Also in a preferred embodiment, the method further comprises the initial steps of:

at the primary site responsive to initiation of a start copy operation, ascertaining at each of the subsystems in said first plurality of subsystems, the subset of the plurality of subsystems forming a copyset group, designating one of the plurality of subsystems as a clocking and checkpoint signal source, each checkpoint signal including a sequence clock value and an increased sequence number;
 at the secondary site responsive to initiation of a start copy operation, ascertaining at each of the subsystems in said second plurality of subsystems, the subset of the plurality of subsystems forming a copyset group, and designating one of the plurality of subsystems as checkpoint signal rendezvous point for synchronising the second plurality of subsystems;
 and wherein the step of sending a checkpoint signal to each of the first plurality of subsystems comprises the steps of periodically generating clocking signals and checkpoint signals by said designated subsystem at the primary site and broadcasting said clocking and checkpoint signals as they occur to other subsystems forming the copyset group at the primary site including itself;
 and wherein the step of sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises for each subsystem in the copyset group, asynchronously forming a local sequence of updated records, embedding said clocking and checkpoint signals into the sequence to form a time discrimi-

nated total ordering of updated records, and sending at least a portion of the sequence to the counterpart subsystem at the secondary site; and wherein the step of coordinating the writing of data updates and checkpoint signals at the secondary site comprises the steps of sending the checkpoint signal to the designated subsystem at the secondary site by each subsystem in the copyset group, and writing the sequences or portions thereof buffered in said subsystems only upon a signal from said designated subsystem indicative of its receipt of all the checkpoint signals sent to it.

Each checkpoint signal signifies that all data updates with a clock signal with a sequence number that is less than the checkpoint signal sequence clock value have been transmitted to the counterpart secondary subsystem.

The invention further provides an asynchronous remote data copy system including a primary site having a first plurality of storage subsystems interconnected by a first coupling means, and a secondary site remote from the primary site having a second plurality of storage subsystems interconnected by a second coupling means, each of the second plurality of storage subsystems being independently coupled to a counterpart one of the first plurality of storage subsystems,

the first plurality of subsystems including means for sending a checkpoint signal to each of the first plurality of subsystems; and means for sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and

the second plurality of subsystems including means for receiving and buffering the data updates and checkpoint signals; and means for coordinating the writing of the data updates based upon the checkpoint signals such that update sequence integrity is preserved at the secondary site.

Preferably, the means for sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises means, responsive to the means for sending a checkpoint signal to each of the first plurality of subsystems, for inserting the checkpoint signal into a data update sequence from each of said first plurality of subsystems to its counterpart subsystem.

Thus the above method and means may be used for the remote copying of data at a secondary DASD subsystems. Typically in such an implementation, the secondary DASD subsystems are peer coupled to counterpart ones of a primary host attached DASD subsystems. A start copy operation is initiated at the host by a message broadcast to all primary DASD subsystems. Each primary builds a configuration table using a local area network or other suitable means as a set associ-

tive device for table building. This table identifies all the primary subsystem participants. Also, each primary synchronizes its local clock or other time reference to that of a designated primary subsystem, the designated primary being operative as a clocking and checkpoint message source. The designated primary periodically sends checkpoint messages having a sequence time value and an increasing checkpoint sequence number, to each primary subsystem. These are logically inserted into its local copy write record sequence. At the secondary host attached DASD subsystems, one of the secondary subsystems is designated as a local synch source. Each secondary subsystem builds a configuration table of copy active subsystems, and couples to the counterpart primary subsystem. Next, each secondary subsystem asynchronously receives and locally buffers a copy sequence from its primary counterpart. Each received sequence includes checkpoint messages embedded therein at the primary after all time stamped write updated records have been sent. Responsive to receipt of a checkpoint message, the secondary subsystem will remit it to the local synch source. A rendezvous is executed by this source over all the checkpoint messages. Thus, each secondary subsystem writes from buffer to DASD upon receipt of the rendezvous completion from the synch source. If one or more of the primary DASD subsystems become unavailable, then their counterparts at the secondary site resign from the copygroup. Such occurs only after completion of any updates in progress.

This approach allows for distributed non-central-system operated control of a sequence-consistent real-time asynchronous data copy from a set of DASD subsystems at a primary location to a set of DASD subsystems at a secondary location, with primary subsystems separately and independently connected to secondary subsystems. This enables update sequence integrity of a data copy at a plurality of subsystems remote from a source of asynchronously independently generated sequence of write operations, there being a first plurality of subsystems at a primary site, the first plurality of subsystems being interconnected by a first coupling means, and a second plurality of subsystems at a site remote from the primary site, the second plurality of subsystems being interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to one of the first plurality of subsystems.

Thus in a system for remote copying of data at a secondary site having a coupling plurality of storage device (DAS) subsystems that are interconnected by a first coupling means, said secondary DASD subsystems being coupled to counterpart ones of a plurality of DASD subsystems at a remote primary site, said secondary DASD subsystems being interconnected by a second coupling means, said system including means at the primary site for initiating a start copy operation, a method is provided comprising the steps of:

- (a) at the primary and secondary sites, forming m copyset groups of DASD subsystems respectively;
 (b) at the primary site:

- (1) causing each of the m subsystems to asynchronously generate a sequence of updated write records;
- (2) ordering each of said sequences by embedding common clock values and a periodic checkpoint message with a common clock value and increasing a sequence number in each of said sequences,
- (3) coupling counterpart DASD subsystems between the sites and asynchronously communicating sequences from each of the primary site subsystems into a buffered portion of a counterpart secondary site subsystem; and

- (c) at the secondary site, writing the buffered sequences to DASD in the counterpart subsystems as a function of each checkpoint message.

An embodiment of the invention will now be described in detail by way of example only, with reference to the following drawings:

Figure 1 is a conventional remote data copy system configuration;
 Figure 2 is remote dual copy system configured in accordance with the present invention;
 Figure 3 is a flow chart showing the general operation of the remote dual copying of Figure 2; and
 Figure 3A-3D are more detailed flow charts showing the operation of the remote dual copying system of Figure 2.

To perform asynchronous remote copying, (1) the sequence of data updates must be determinable at a local site; (2) that sequence must be communicable by the local site to a remote site; and (3) the remote site must be able to use the sequence to control the updating at the remote site.

As has been before mentioned, prior art asynchronous copy systems which include multiple subsystem require a central system for providing the appropriate sequences of data. However, system configurations may be found in which it is not desired to pass update data between the primary and secondary locations through a centralized communication service. Rather in those configurations it is desired to directly connect DASD subsystems at the primary and secondary sites via independent subsystem-to-subsystem communication links. Such a system 10 is shown in Figure 1. The system 10 includes a host 11 which provides data to primary subsystems 12'. As is shown each of the links between a primary DASD subsystem 12' and its peer coupled secondary DASD subsystem 14' is independent. As a result, this type of system would inherently be in-

capable of write operations that are sequentially consistent.

To more specifically describe the problem, consider as an example a sequence of three writes from a conventional database management system (DBMS) as it is about to commit a transaction. The example is representative of an information management service (IMS) system:

1. The DBMS writes to its log data set; the record written contains old data base (DB) data, new DB data (that is being changed by this transaction), and a record of its intent to commit (finalize) this transaction.
2. The DBMS waits for a DASD I/O operation to report that it has completed, then it updates its data base data sets, which are different volumes that are on a different DASD subsystems. This writing of the new DB data overwrites and thus destroys the old DB records.
3. The DBMS waits for the second DASD I/O operation to be posted complete, then it writes a commit record to its log data set on the first volume. This commit record 'guarantees' to future IMS recovery processes that the data base data set (DASD volumes) have been updated.

Now consider operation of an asynchronous remote copy system using a multiple subsystem configuration such as illustrated in Figure 1. In this embodiment, the DBMS log data set is on a volume configured to the top-most DASD subsystem pair (DASD 1 and DASD 1') and the data base volumes are on the DASD subsystem pair shown second from the top, (DASD 2 and DASD 2'). In this example, the primary subsystem 1 is lightly loaded, thus it processes its queued work with no delay, while the primary subsystem 2 is heavily loaded and is experiencing some delay in processing its queued work. Queued work includes the forwarding of updated data to its remote copy peer subsystem.

The following sequence would describe the operation in the present example:

1. Write I/O (1) is completed from application to subsystem 1.
2. Write I/O (2) is completed from application to subsystem 2.
3. Primary Subsystem 1 sends data for I/O (1) to secondary subsystem 1, which applies it to its cache memory copy of the DASD volume.
4. Write I/O (3) is completed from application system to subsystem 1.
5. Primary Subsystem 1 sends data for I/O (3) to secondary subsystem 1, which applies it to its cache memory copy of the DASD volume.
6. Primary Subsystem 2 sends data for I/O (2) to secondary subsystem 2, which applies to its cache memory copy of the DASD volume.

If a primary site failure occurs after step 5 and before step 6 there would be corrupted data copied into the system. Since the failure rendered primary subsystems inaccessible, the data from I/O (2) will not be at the secondary site.

The essence of data sequence consistency in a remote copy service is to ensure at such a time as the remote DASD must be used for recovery, that the new data from second operation I/O above will only be seen if the data from the first I/O is also present, and that the data from the third I/O will only be present if the second I/O is also present. Consider the data integrity loss if the third I/O were present but if the second I/O was not. The DBMS log would tell recovery processes that the data base would receive valid data. This would either result in a DBMS failure or business application error.

The sequence of I/O operations at the primary subsystems is DASD 1, DASD 2, then back to DASD 1; but because of the load on DASD subsystem 2, it is delayed in sending its I/O (2) such that it has not arrived by the time I/O (3) from DASD subsystem 1 was received by secondary DASD subsystem 1'. With no control of transmission subsystem data to copy volumes, subsystem 1 would update its copy volume with both 1 and 3. If at that time a disaster befell that primary system and operations were directed to resume at the secondary site (such contingency being the reason for having a real-time remote DASD copy), the recovering DBMS would find a log record that said that data base records were written (in I/O 2) while the data for I/O 2 would not be on the DASD of secondary subsystem 2.

With such independent links, determination of sequence of writes among independent primary DASD subsystems, communication of that information to the set of subsystems at the secondary location, and control of the sequence of updates among the independent DASD subsystems at the secondary has been a problem heretofore. However, as explained in detail below, a system and method are provided for sequence identification, communication, and update control that permit sequence-consistent remote DASD data copy from multiple independent DASD subsystems at one location to a set of secondary DASD subsystems, each subsystem being independently connected to peer subsystems at the first location.

Figure 2 is a remote dual copy system 20 including a host 11 which provides data to primary subsystems 12. The system 20 includes a first group of DASD subsystems 12 which are located at a primary site and a second group of DASD subsystems 14 which are located at a site that is remote from the primary site. The system 20 includes couplers 16 and 18 which provide for interconnections of the DASD subsystems 12 and 14, respectively.

Referring now to Figure 2, system 20 for achieving update sequence integrity without a centralized communication system is based on the presence of the following configurations.

1. Multiple primary location DASD subsystems each interconnected via one or more communication links to a peer DASD subsystem at the secondary location.

2. A coupling connection 16 of all subsystems at the primary site, and a similar coupling connection 18 of all subsystems at the secondary site. While not shown, one of ordinary skill in the art will readily recognize multiple physical connections may be incorporated for connection redundancy. The term "coupling" is used for convenience; any suitable physical network interconnection means may be used such as a local area network (LAN) or the like.

3. Each subsystem has a 'clock' or similar synchronizing signal process that can be synchronized with a value from another subsystem, communicated via the coupling 16.

Three steps are utilized to achieve copy update sequence integrity: (1) determination of sequence of write operations among all DASD subsystems at the primary; (2) communication of that sequence information to the secondary; and (3) use of that information by the secondary DASD subsystems to control the sequence of update writes across all secondary DASD. These three steps are described in more detail hereinbelow for a copy system using LAN interconnections at the secondary and primary subsystems.

1. Use of the LAN interconnection among the primary DASDs to distribute a sequencing signal, such as a time clock, to all subsystems is used to associate a sequence/time value with each DASD write of data to be copied to a secondary DASD subsystem (and the sending of that value along with update data to the secondary);

2. Propagation of periodic synchronizing-time-denominated checkpoint signals among the primary DASD subsystems that in turn are communicated by each primary subsystem to its peer-connected secondary subsystem(s); and

3. Use of the LAN interconnection among secondary DASD subsystems to coordinate their DASD update writing of copy data received from primary subsystems.

Referring now to Figure 3, what is shown is a flow chart of the general operation 100 of such a system that is located within the primary subsystem. First, a start copy operation is sent to all the primary subsystems to activate communication between primary subsystems, via step 102. Then individual subsystems provide the appropriate sequence information to all the copy active primary subsystems, via step 104. Thereafter the primary subsystems communicate the sequence information to the peer coupled secondary subsystems, via step 106. Finally, that sequence information is utilized to control secondary subsystem updates via step 108.

These steps are described in detail below:

Start copy operation (step 102)

Remote copy operations for each DASD subsystem must be started by some instruction to that subsystem. The instruction may come via command from a host system (the same system as may be executing applications that write to DASD), or it may be provided via subsystem-local interface such as from a subsystem operator console or similar means.

Description of DASD write sequence at the primary (step 104)

Refer now to Figure 3A and the following discussion. Irrespective of whether the command comes from the host system or from subsystem local interface, the start copy instruction identifies the DASD to be copied and causes the subsystem to activate communications with secondary subsystem(s) to which it is connected.

The start copy operation also activates communication from that DASD subsystem to other primary DASD subsystems (step 202). The LAN connection addresses of other primary DASD subsystems may be configured to each subsystem or it may be incorporated in information contained in the start copy instruction to the subsystem. When copy is started at a subsystem, this subsystem sends a subsystem copy active message to all other primary subsystems it has addresses for via step 204. All primary subsystems build and maintain configuration tables such that each subsystem knows the identity of all primary subsystems participating in the copy process via step 206. Each subsystem receiving an copy active message from another subsystem marks that subsystem as a copy active in its configuration tables.

As a part of exchanging copy active messages with the other primary systems, the subsystems synchronize their sequence clock processes and select one subsystem to be a master source for a timing value synchronization signal via step 208. This is a clock synchronization process, not described here since such processes are well known in the art. Note that clock synchronization must be able to maintain clock drift such that maximum drift is substantially less than the time for a subsystem to receive a write command from a system, perform a write to cache, signal end of I/O operation, and for the host to process the I/O completion status and start a new DASD I/O write operation.

As write operations are performed by each subsystem, in a preferred embodiment, the subsystem includes the then current time sequence signal value with other control information that is sent with the DASD data to its connected secondary subsystem. The primary system DASD write I/O operation continues without delay, extended only by the time necessary to construct control information for the data to be sent to the secondary.

DASD and control data are buffered in the secondary subsystem on receipt. Update of secondary DASD copy is deferred until released by secondary sequence control (described below).

Communication of sequence information to secondary DASD subsystems (step 106)

Refer now to Figure 3B and the following discussion. The subsystem that is providing the time synchronizing signal source will periodically send a checkpoint message to all primary copy-active subsystems via step 302. This may be included with the time sync signal or sent separately as appropriate for the local interconnection protocol used. The checkpoint message includes a sequence time value and a checkpoint sequence number that is incremented by one for each checkpoint message. Each subsystem on receiving the checkpoint communication will logically insert it in its transmission stream to the secondary subsystem(s) to which it is connected, sending it to secondary subsystem(s) only after all DASD and control information with an earlier sequence time signal value have been sent via step 304.

Use of sequence information by secondary DASD subsystems to control secondary DASD updates (step 108)

Refer now to Figures 3C and 3D and the following discussion. Copy operations in secondary systems are activated by copy active messages from their connected primary subsystems via step 402. (An enabling startup instruction from a system or local console at the secondary may also be required. That aspect of copy operation and control is performed in a conventional manner.) Secondary subsystems build and maintain copy active tables such that each subsystem has data relating to the identity of all secondary subsystems participating in the copy process, via step 404. A synchronizing control master is selected from among the secondary subsystems, using the local interconnect (eg a LAN) similar to the manner in which a primary synchronizing signal source was selected, via step 406. Such distributed control schemes are well known and need not be described here.

Secondary subsystems receive and buffer DASD data and associated control information from their connected primary subsystems. This received data and control information is logically grouped and sequenced by the primary synchronizing signal value. For maximum protection, received data and control information buffering should be in a non-volatile storage medium.

At some point, each subsystem will receive a checkpoint control message via step 408 that signifies that all DASD update data and control with a primary synchronizing signal value equal to or less than the checkpoint time sync value has been sent to that secondary subsystem. A checkpoint message, when received, is sent

by the receiving subsystem to the secondary master subsystem, which performs a rendezvous for that checkpoint message from all copy-active secondary subsystems, including itself, via step 410.

When the rendezvous is complete via step 412 for a given checkpoint value, the secondary master subsystem sends a message to all secondary copy-active subsystems to release update data up to the primary sequence time value of the checkpoint. Each subsystem then marks itself in an internal update in progress state and applies the updates to its secondary DASD copy via step 414. (Note: The actual process of applying the buffered DASD copy data may require only adjusting cache directory entries.)

The update in progress state must be maintained through subsystem resets and power off along with buffered data and control info from the primary (and other copy service state information). It is used as a must-complete type operation control that precludes any host system access to secondary DASD when that state is active. This ensures that an application takeover at the secondary cannot see partial updated DASD data. That is, it can not access a sequence-inconsistent DASD copy. Update for the checkpoint interval must be complete before a user can access the copy DASD subsystems records.

When the updating of DASD copy data has completed, the subsystem sends an update complete for checkpoint message (identifying the specific checkpoint) for that checkpoint to the secondary master, via step 416. When update complete signal for checkpoint messages have been received at the master from all subsystems including the master, the master then sends a reset update in progress state message to all secondary subsystems to allow secondary copy data to again be accessible to attached systems, via step 418.

In a variation to the preferred embodiment, if the coupling means among all primary subsystems can reliably propagate every checkpoint message to all subsystems in substantially less time than the time for an I/O operation cycle then the preceding processes could function without a clock and clock sync. The arrival time of a checkpoint message at each primary subsystem would be precise enough to define update sequence. All updates within a checkpoint would be considered to have occurred at the same time.

The steps of subsystem operation for the three I/Os described previously, using the approach described above is discussed herein below.

1. DASD 1, 2 and 3, and 4 exchange 'copy active' messages. Primary subsystem 3 has become the master source for timing value sync signal. Secondary system 4 has become the secondary master for rendezvous of checkpoint messages.

2. Write I/O (1) is completed from application to subsystem 1 at time 'a'.

3. Write I/O (2) is completed from application to sub-

system 2 at time 'b'.

4. Primary Subsystem 1 sends data for I/O (1) to secondary subsystem 1 along with its associated time value 'a'. Secondary subsystem 1 buffers the data but does not apply it to its cache copy for the DASD volume.

5. Subsystem 3 sends a checkpoint message containing checkpoint sequence number 'n' and time value 'b'.

6. Write I/O (3) is completed from application system to subsystem 1 at time 'c'.

7. Primary Subsystem 1 sends data for I/O (3) to secondary subsystem 1 along with its associated time value 'c'. Secondary subsystem 1 buffers the data but does not apply it to its cache copy for the DASD volume.

8. Primary subsystem 1 receives and processes the checkpoint message sent by subsystem 3 in step 5.

9. Primary subsystem 1 sends checkpoint message to secondary subsystem 1, which forwards it to secondary subsystem 4.

10. Primary Subsystem 2 sends data for I/O (2) to secondary subsystem 2 along with its associated time value 'b'. Secondary subsystem 2 buffers the data but does not apply it to its cache copy for the DASD volume.

11. Primary subsystem 2 receives and processes the checkpoint message sent by subsystem 3 in step 5.

12. Primary subsystem 2 sends checkpoint message to secondary subsystem 2, which forwards it on to secondary subsystem 4.

13. At some point between steps 5 and 13, subsystems 3 and 4 have sent the checkpoint message 'n' to their secondary subsystems. Secondary subsystem 3 has forwarded it to secondary subsystem 4.

14. Secondary subsystem 4 sends a 'release' message to secondary subsystems 1, 2 and 3.

15. Secondary subsystem 3 having no update immediately returns an "update complete" message to secondary subsystem 4.

16. Secondary subsystem 1 enters 'update in progress' state, then applies update (1). It does not apply update (3) since its sync time value 'c' is greater than checkpoint time value 'b'. It then sends an update complete message to secondary subsystem 4.

17. Secondary subsystem 2 enters update state and applies update (2), and sends an update complete message to secondary subsystem 4.

18. Secondary subsystem 4 having received update complete messages from all secondary subsystems, sends a 'reset update in progress state' message to secondary subsystems 1, 2 and 3.

Now if a primary site failure happens at any point in the above sequence the secondary DASD will either show none of the updates, or will show updates (1) and

(2). Update (3) will not be 'applied' to secondary subsystem 1's DASD and cache until the next checkpoint sequence.

Accordingly, through the present system a sequence consistent real-time asynchronous copy system is provided that does not require the use of central communications service. In so doing, a system is provided that requires minimal modification, while utilizing existing capabilities within the DASD subsystems to provide for sequence modifications.

Claims

1. A method of operating an asynchronous remote data copy system (20) including a primary site having a first plurality of storage subsystems (12) interconnected by a first coupling means (16), and a secondary site remote from the primary site having a second plurality of storage subsystems (14) interconnected by a second coupling means, each of the second plurality of storage subsystems being independently coupled to a counterpart one of the first plurality of storage subsystems, said method comprising the steps of:

in the first plurality of subsystems;
 sending (302) a checkpoint signal to each of the first plurality of subsystems;
 sending (304) data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and in the second plurality of subsystems;
 receiving (408) and buffering the data updates and checkpoint signals;
 and coordinating the writing (414) of the data updates based upon the checkpoint signals such that update sequence integrity is preserved at the secondary site.

2. The method of claim 1, wherein the step of sending a checkpoint signal to each of the first plurality of subsystems further comprises the step of distributing a sequence signal at the primary site, said checkpoint signal including information indicative of the value of the sequence signal.

3. The method of claim 2, wherein said step of sending data updates and a checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises the steps of:

generating asynchronously a sequence of data updates at each of said first plurality of subsystems;
 ordering (108) each of said sequences in accordance with the distributed sequence signal and checkpoint signal; and

asynchronously communicating sequences of updates from each of the first plurality of subsystems to a counterpart subsystem.

4. The method of any preceding claim further comprising the initial steps at the primary site of:

activating (202) each of the first plurality of subsystems to communicate with the other subsystems in said first plurality of subsystems;
 activating each of the first plurality of subsystems to communicate with its counterpart coupled subsystem in the second plurality of subsystems;
 building (206) at least one configuration table in each of the first plurality of subsystems such that each of the first plurality of subsystems can identify all of the other subsystems in said first plurality of subsystems; and
 synchronizing (208) the first plurality of subsystems.

5. The method of any preceding claim, further comprising the initial steps at the secondary site of:

receiving (402) copy active messages in the second plurality of subsystems;
 building (404) copy active tables in the second plurality of subsystems; and
 synchronizing (406) copy operations of the second plurality of subsystems;
 and wherein the coordinating step further comprises the step of performing a rendezvous (410) for all checkpoint signals in the second plurality of subsystems.

6. The method of claim 1, further comprising the initial steps of:

at the primary site responsive to initiation of a start copy operation, ascertaining at each of the subsystems in said first plurality of subsystems, the subset of the plurality of subsystems forming a copyset group, and designating one of the plurality of subsystems as a clocking and checkpoint signal source, each checkpoint signal including a sequence clock value and an increased sequence number;
 at the secondary site responsive to initiation of a start copy operation, ascertaining at each of the subsystem in said second plurality of subsystems, the subset of the plurality of subsystems forming a copyset group, and designating one of the plurality of subsystems as checkpoint signal rendez-vous point for synchronizing the second plurality of subsystems;
 and wherein the step of sending a checkpoint signal to each of the first plurality of subsystems

comprises the steps of periodically generating clocking signals and checkpoint signals by said designated subsystem at the primary site and broadcasting said clocking and checkpoint signals as they occur to other subsystems forming the copyset group at the primary site including itself;

and wherein the step of sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises for each subsystem in the copyset group, asynchronously forming a local sequence of updated records, embedding said clocking and checkpoint signals into the sequence to form a time discriminated total ordering of updated records, and sending at least a portion of the sequence to the counterpart subsystem at the secondary site;

and wherein the step of coordinating the writing of data updates and checkpoint signals at the secondary site comprises the steps of sending (410) the checkpoint signal to the designated subsystem at the secondary site by each subsystem in the copyset group, and writing the sequences or portions thereof buffered in said subsystems only upon a signal from said designated subsystem indicative of its receipt of all the checkpoint signals sent to it (412).

7. The method according to claim 6, wherein each checkpoint signal signifies that all data updates with a clock signal with a sequence number that is less than the checkpoint signal sequence clock value have been transmitted to the counterpart secondary subsystem.

8. An asynchronous remote data copy system (20) including a primary site having a first plurality of storage subsystems (12) interconnected by a first coupling means (16), and a secondary site remote from the primary site having a second plurality of storage subsystems (14) interconnected by a second coupling means, each of the second plurality of storage subsystems being independently coupled to a counterpart one of the first plurality of storage subsystems,

the first plurality of subsystems including means (12) for sending a checkpoint signal to each of the first plurality of subsystems; and means (12) for sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and the second plurality of subsystems including means (14) for receiving and buffering the data updates and checkpoint signals; and means (14) for coordinating the writing of the data updates based upon the checkpoint signals such

that update sequence integrity is preserved at the secondary site.

9. The system of claim 8, wherein the means for sending a checkpoint signal to each of the first plurality of subsystems further includes means for distributing a sequence signal at the primary site, said checkpoint signal including information indicative of the value of the sequence signal.
10. The system of claim 8 or 9 in which the means for sending data updates and the checkpoint signal to each of the counterpart coupled second plurality of subsystems comprises means, responsive to the means for sending a checkpoint signal to each of the first plurality of subsystems, for inserting the checkpoint signal into a data update sequence from each of said first plurality of subsystems to its counterpart subsystem.

Patentansprüche

1. Verfahren zum Betreiben eines asynchronen Datenfernkopiersystems (20), das eine Primärstation enthält, die eine erste Vielzahl von Speicher-Subsystemen (12) hat, die durch ein erstes Kopplungsmittel (16) miteinander verbunden sind, und eine von der Primärstation entfernte Sekundärstation, die eine zweite Vielzahl von Speicher-Subsystemen (14) hat, die durch ein zweites Kopplungsmittel miteinander verbunden sind, wobei jedes der zweiten Vielzahl der Speicher-Subsysteme unabhängig von den anderen mit einem Gegenstück der ersten Vielzahl der Speicher-Subsysteme verbunden ist, wobei das Verfahren die folgenden Schritte umfaßt:

in der ersten Vielzahl der Subsysteme

Senden (302) eines Prüfpunktsignals an jedes der ersten Vielzahl der Subsysteme;

Senden (304) von Datenaktualisierungen und des Prüfpunktsignals an jedes der mit einem Gegenstück verbundenen zweiten Vielzahl der Subsysteme; und

in der zweiten Vielzahl der Subsysteme

Empfangen (408) und Puffern der Datenaktualisierungen und der Prüfpunktsignale;

und Koordinieren des Schreibens (414) der Datenaktualisierungen auf der Grundlage der Prüfpunktsignale, so daß die Integrität der Aktualisierungsfolgen an der Sekundärstation gewahrt bleibt.

2. Verfahren nach Anspruch 1, wobei der Schritt des Sendens eines Prüfpunktsignals an jedes der ersten Vielzahl der Subsysteme des weiteren den Schritt der Verteilung eines Folgesignals an der Primärstation umfaßt, wobei das Prüfpunktsignal Informationen enthält, die den Wert des Folgesignals anzeigen. 5
3. Verfahren nach Anspruch 2, wobei der Schritt des Sendens von Datenaktualisierungen und eines Prüfpunktsignals an jedes der mit einem Gegenstück verbundenen zweiten Vielzahl der Subsysteme die folgenden Schritte umfaßt: 10
- Asynchrones Erzeugen einer Folge von Datenaktualisierungen an jedem der ersten Vielzahl der Subsysteme; 15
- Ordnen (108) einer jeden der Folgen entsprechend dem verteilten Folgesignal und dem Prüfpunktsignal; und 20
- Asynchrones Übertragen von Aktualisierungsfolgen von jedem der ersten Vielzahl der Subsysteme an ein Subsystem-Gegenstück. 25
4. Verfahren nach einem vorhergehenden Anspruch, das an der Primärstation des weiteren die folgenden ersten Schritte umfaßt: 30
- Aktivieren (202) eines jeden der ersten Vielzahl der Subsysteme, damit es mit den anderen Subsystemen in der ersten Vielzahl der Subsysteme kommuniziert; 35
- Aktivieren eines jeden der ersten Vielzahl der Subsysteme, damit es mit seinem Subsystem-Gegenstück in der zweiten Vielzahl der Subsysteme, mit dem es verbunden ist, kommuniziert; 40
- Erstellen (206) von mindestens einer Konfigurationstabelle in jedem der ersten Vielzahl der Subsysteme, so daß jedes der ersten Vielzahl der Subsysteme alle anderen Subsysteme in der ersten Vielzahl der Subsysteme erkennen kann; und 45
- Synchronisieren (208) der ersten Vielzahl der Subsysteme. 50
5. Verfahren nach einem vorhergehenden Anspruch, das an der Sekundärstation des weiteren die folgenden ersten Schritte umfaßt: 55
- Empfangen (402) von Kopiervorgang-Aktiv-Nachrichten in der zweiten Vielzahl der Subsysteme;
- Erstellen (404) von Kopiervorgang-Aktiv-Tabellen in der zweiten Vielzahl der Subsysteme; und
- Synchronisieren (406) von Kopieroperationen der zweiten Vielzahl der Subsysteme;
- und wobei der Koordinierungsschritt des weiteren den Schritt der Durchführung eines Treffens (410) aller Prüfpunktsignale in der zweiten Vielzahl der Subsysteme umfaßt.
6. Verfahren nach Anspruch 1, das des weiteren die folgenden ersten Schritte umfaßt:
- an der Primärstation als Antwort auf die Einleitung einer Startkopieroperation an jedem der Subsysteme in der ersten Vielzahl der Subsysteme Feststellen der Untergruppe der Vielzahl der Subsysteme, die eine Kopiersatzgruppe bildet, und Bestimmen eines der Vielzahl der Subsysteme zu einer Takt- und Prüfpunktsignalquelle, wobei jedes Prüfpunktsignal einen Folgetaktwert und eine erhöhte Folgenummer enthält;
- an der Sekundärstation als Antwort auf die Einleitung einer Startkopieroperation an jedem der Subsysteme in der zweiten Vielzahl der Subsysteme Feststellen der Untergruppe der Vielzahl der Subsysteme, die eine Kopiersatzgruppe bildet, und Bestimmen eines der Vielzahl der Subsysteme zu einem Prüfpunktsignal-Treffpunkt, um die zweite Vielzahl der Subsysteme zu synchronisieren;
- und wobei der Schritt des Sendens eines Prüfpunktsignals an jedes der ersten Vielzahl der Subsysteme die Schritte der periodischen Erzeugung von Taktsignalen und Prüfpunktsignalen durch das ausgewählte Subsystem an der Primärstation und die Übertragung der Takt- und Prüfpunktsignale in der Folge ihres Auftretens an andere Subsysteme, welche die Kopiersatzgruppe an der Primärstation bilden, einschließlich ihrer selbst, umfaßt;
- und wobei der Schritt des Sendens von Datenaktualisierungen und des Prüfpunktsignals an jedes der mit einem Gegenstück verbundenen zweiten Vielzahl der Subsysteme für jedes Subsystem in der Kopiersatzgruppe die asynchrone Bildung einer lokalen Folge von aktualisierten Datensätzen umfaßt, wobei die Takt- und Prüfpunktsignale in die Folge eingebunden werden, um eine nach dem Zeitpunkt unterschiedene Gesamtanordnung von aktualisierten Datensätzen zu bilden, und Senden von

mindestens einem Teil der Folge an das Subsystem-Gegenstück an der Sekundärstation;

und wobei der Schritt der Koordination des Schreibens von Datenaktualisierungen und Prüfpunktsignalen an der Sekundärstation die Schritte des Sendens (410) des Prüfpunktsignals an das ausgewählte Subsystem an der Sekundärstation durch jedes Subsystem in der Kopiersatzgruppe und das Schreiben der Folgen oder von Teilen der Folgen, die in den Subsystemen gepuffert sind, erst nach einem Signal von dem ausgewählten Subsystem, das dessen Empfang aller an das Subsystem gesandten Prüfpunktsignale (412) anzeigt, umfaßt.

7. Verfahren nach Anspruch 6, wobei jedes Prüfpunktsignal anzeigt, daß alle Datenaktualisierungen mit einem Taktsignal, dessen Folgennummer kleiner als der Prüfpunktsignalfolgetaktwert ist, an das sekundäre Subsystem-Gegenstück übertragen worden sind.

8. Asynchrones Datenfernkopiersystem (20), das eine Primärstation enthält, die eine erste Vielzahl von Speicher-Subsystemen (12) hat, die durch ein erstes Kopplungsmittel (16) miteinander verbunden sind, und eine von der Primärstation entfernte Sekundärstation, die eine zweite Vielzahl von Speicher-Subsystemen (14) hat, die durch ein zweites Kopplungsmittel miteinander verbunden sind, wobei jedes der zweiten Vielzahl der Speicher-Subsysteme unabhängig von den anderen mit einem Gegenstück der ersten Vielzahl der Speicher-Subsysteme verbunden ist,

wobei die erste Vielzahl der Subsysteme ein Mittel (12) enthält, um an jedes der ersten Vielzahl der Subsysteme ein Prüfpunktsignal zu senden; und ein Mittel (12), um Datenaktualisierungen und das Prüfpunktsignal an jedes der mit einem Gegenstück verbundenen zweiten Vielzahl der Subsysteme zu senden; und

die zweite Vielzahl der Subsysteme ein Mittel (14) enthält, um die Datenaktualisierungen und die Prüfpunktsignale zu empfangen und zu puffern; und ein Mittel (14), um das Schreiben der Datenaktualisierungen auf der Grundlage der Prüfpunktsignale zu koordinieren, so daß die Integrität der Aktualisierungsfolgen an der Sekundärstation gewahrt bleibt.

9. System nach Anspruch 8, wobei das Mittel, das dazu dient, ein Prüfpunktsignal an jedes der ersten Vielzahl der Subsysteme zu senden, des weiteren ein Mittel enthält, um ein Folgesignal an der Primär-

station zu verteilen, wobei das Prüfpunktsignal Informationen enthält, die den Wert des Folgesignals anzeigen.

10. System nach Anspruch 8 oder 9, in dem das Mittel, das dazu dient, Datenaktualisierungen und das Prüfpunktsignal an jedes der mit einem Gegenstück verbundenen zweiten Vielzahl der Subsysteme zu senden, ein Mittel umfaßt, das auf das Mittel zum Senden eines Prüfpunktsignals an jedes der ersten Vielzahl der Subsysteme anspricht, um das Prüfpunktsignal in eine Datenaktualisierungsfolge von jedem der ersten Vielzahl der Subsysteme an dessen Subsystem-Gegenstück einzufügen.

Revendications

1. Procédé d'exploitation d'un système de réplication asynchrone et à distance des données (20), comprenant un premier site ayant une première pluralité de sous-systèmes de mémoire (12) interconnectés par un premier moyen de couplage (16), et un site secondaire, distant du premier site, ayant une deuxième pluralité de sous-systèmes de mémoire (14) interconnectés par un deuxième moyen de couplage, chacun des sous-systèmes de la deuxième pluralité de sous-systèmes de mémoire étant couplé indépendamment à son pendant dans la première pluralité de sous-systèmes de mémoire, ledit procédé comprenant les phases qui consistent à :

dans la première pluralité de sous-systèmes :
envoyer (302) un signal de point de contrôle à chacun des sous-systèmes de la première pluralité de sous-systèmes ;
envoyer (304) des mises à jour de données et le signal de point de contrôle à chacun des pendants de la deuxième pluralité de sous-systèmes couplés ; et

dans la deuxième pluralité de sous-systèmes :
recevoir (408) et enregistrer dans une mémoire tampon les mises à jours des données et les signaux de point de contrôle ;
et coordonner l'écriture (414) des mises à jour des données en se basant sur les signaux de point de contrôle pour que l'intégrité de la séquence de mise à jour soit préservée sur le site secondaire.

2. Procédé selon la revendication 1, où la phase d'envoi d'un signal point de contrôle à chacun des sous-systèmes de la pluralité de sous-systèmes comprend une phase qui consiste à répartir un signal de séquence sur le premier site, ledit signal de point de contrôle comprenant l'information indi-

quant la valeur du signal de séquence.

3. Procédé selon la revendication 2, où ladite phase d'envoi de mises à jour de données et d'un signal de point de contrôle à chacun des pendants couplés de la deuxième pluralité de sous-systèmes comprend les phases qui consistent à :

engendrer de manière asynchrone une séquence de mises à jour de données dans chacun des sous-systèmes de ladite première pluralité de sous-systèmes ;

classer (108) chacune des dites séquences conformément au signal de séquence réparti et au signal de point de contrôle ; et

transmettre de manière asynchrone les séquences de mises à jour, de chaque système de la première pluralité de sous-systèmes à son sous-système équivalent.

4. Procédé selon l'une quelconque des revendications précédentes, comprenant en outre, sur le premier site, les phases initiales qui consistent à :

activer (202) chacun des sous-systèmes de la première pluralité de sous-systèmes pour communiquer avec les autres sous-systèmes de ladite première pluralité de sous-systèmes ;

activer chacun des sous-systèmes de la première pluralité de sous-systèmes pour qu'il communique avec son pendant dans la deuxième pluralité de sous-systèmes couplés ;

construire (206) au moins un tableau de configuration dans chacun des sous-systèmes de la première pluralité de sous-systèmes pour que chaque sous-système de la première pluralité de sous-systèmes puisse identifier tous les autres sous-systèmes de ladite première pluralité de sous-systèmes ; et

synchroniser (208) la première pluralité de sous-systèmes.

5. Procédé selon l'une quelconque des revendications précédentes, comprenant en outre, sur le deuxième site, les phases initiales suivantes :

recevoir (402) les messages de copies actifs dans la deuxième pluralité de sous-systèmes ;

construire (404) des tables actives de copie dans la deuxième pluralité de sous-systèmes ; et

synchroniser (406) les opérations de copie de la deuxième pluralité de sous-systèmes ;

et où la phase de coordination comprend en outre la phase qui consiste à exécuter le rendez-vous (410) de tous les signaux de contrôle dans la deuxième pluralité de sous-systèmes.

6. Procédé selon la revendication 1, comprenant en outre les phases initiales suivantes :

sur le premier site, en réponse au lancement d'une opération de copie, vérifier chacun des sous-systèmes de ladite première pluralité de sous-systèmes, le sous-ensemble de la pluralité de sous-systèmes formant un groupe de copie, et désigner l'un des éléments de la pluralité de sous-systèmes comme source des signaux de synchronisation et de point de contrôle, chaque signal de point de contrôle comprenant une valeur de synchronisation pour la séquence et un numéro de séquence plus élevé ;

sur le site secondaire, en réponse au lancement d'une opération de copie, vérifier chacun des sous-systèmes de ladite deuxième pluralité de sous-systèmes, le sous-ensemble composé de la pluralité de sous-systèmes formant un groupe de copie, et désigner l'un des éléments de la pluralité de sous-systèmes comme point de rendez-vous des signaux de point de contrôle pour synchroniser la deuxième pluralité de sous-systèmes ;

et où la phase d'envoi d'un signal de point de contrôle à chaque élément de la première pluralité de sous-systèmes comprend les phases qui consistent à faire générer périodiquement, par ledit sous-système désigné sur le site primaire, des signaux de synchronisation et des signaux de point de contrôle et à diffuser les dits signaux de synchronisation et de point de contrôle, une fois créés, vers les autres sous-systèmes du groupe de copie sur le site primaire, y compris vers lui-même ;

et où la phase d'envoi des mises à jour de données et du signal de point de contrôle à chaque équivalent dans la deuxième pluralité de sous-systèmes couplés comprend les étapes suivantes pour chaque sous-système du groupe de copie : former de manière asynchrone une séquence locale avec les enregistrements mis à jour, imbriquer les dits signaux de synchronisation et de point de contrôle dans cette séquence pour former un classement complet distinguant dans le temps les enregistrements mis à jour, et envoyer au moins une partie de

la séquence au sous-système équivalent dans le site secondaire ;

et où la phase qui consiste à coordonner l'écriture des mises à jour des données et les signaux de point de contrôle sur le site secondaire comprend les phases suivantes : chaque sous-système du groupe de copie envoie (410) le signal de point de contrôle au sous-système qui a été désigné dans le site secondaire, et les séquences ou les parties des séquences enregistrées dans les mémoires tampons des dits sous-systèmes ne sont écrites que lorsque ledit sous-système désigné indique par un signal qu'il a reçu tous les signaux de point de contrôle qui lui ont été envoyés (412).

7. Procédé selon la revendication 6, où chaque signal de point de contrôle indique que toutes les mises à jour de données ayant un signal de synchronisation avec un numéro de séquence inférieur à la valeur de synchronisation de séquence du signal de point de contrôle ont été transmises au sous-système secondaire équivalent.
8. Système de réplication de données, asynchrone, à distance (20) incluant un site principal ayant une première pluralité de sous-systèmes de mémoire (12) interconnectés par un premier moyen de couplage (16), et un site secondaire distant du site principal ayant une deuxième pluralité de sous-systèmes de mémoire (14) interconnectés par un deuxième moyen de couplage, chacun des systèmes de la deuxième pluralité de sous-systèmes étant couplé indépendamment à son pendant dans la première pluralité de sous-systèmes de mémoire,

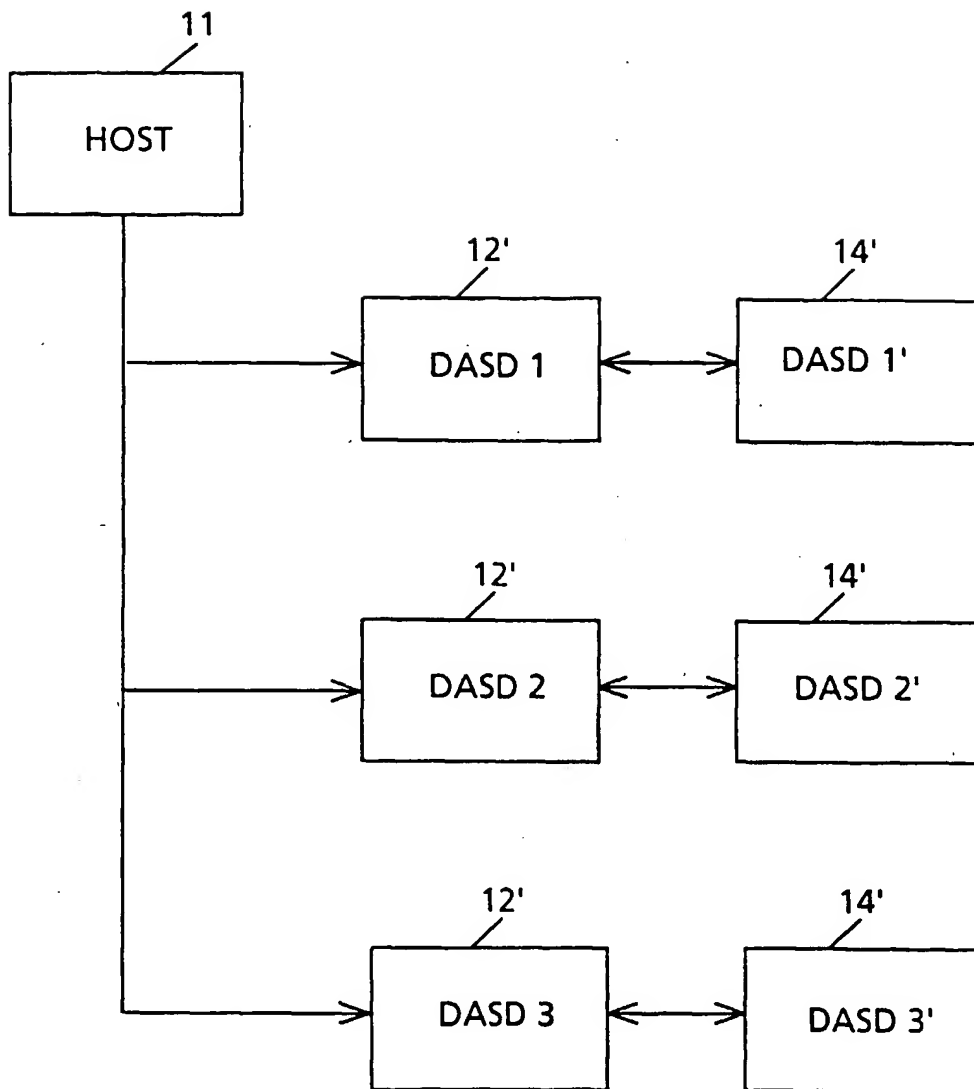
la première pluralité de sous-systèmes comprenant un moyen (12) pour envoyer un signal de point de contrôle à chaque système de la première pluralité de sous-systèmes ; et un moyen (12) pour envoyer les mises à jour des données et le signal de point de contrôle à chaque pendant couplé de la deuxième pluralité de sous-systèmes ; et

la deuxième pluralité de sous-systèmes comprenant un moyen (14) pour recevoir et conserver dans une mémoire tampon les mises à jour des données et les signaux de point de contrôle ; et un moyen (14) pour coordonner l'écriture des mises à jour de données en se basant sur les signaux de point de contrôle de telle sorte que l'intégrité de la séquence de mise à jour soit préservée sur le site secondaire.

9. Système selon la revendication 8, où le moyen pour envoyer un signal de point de contrôle à chaque

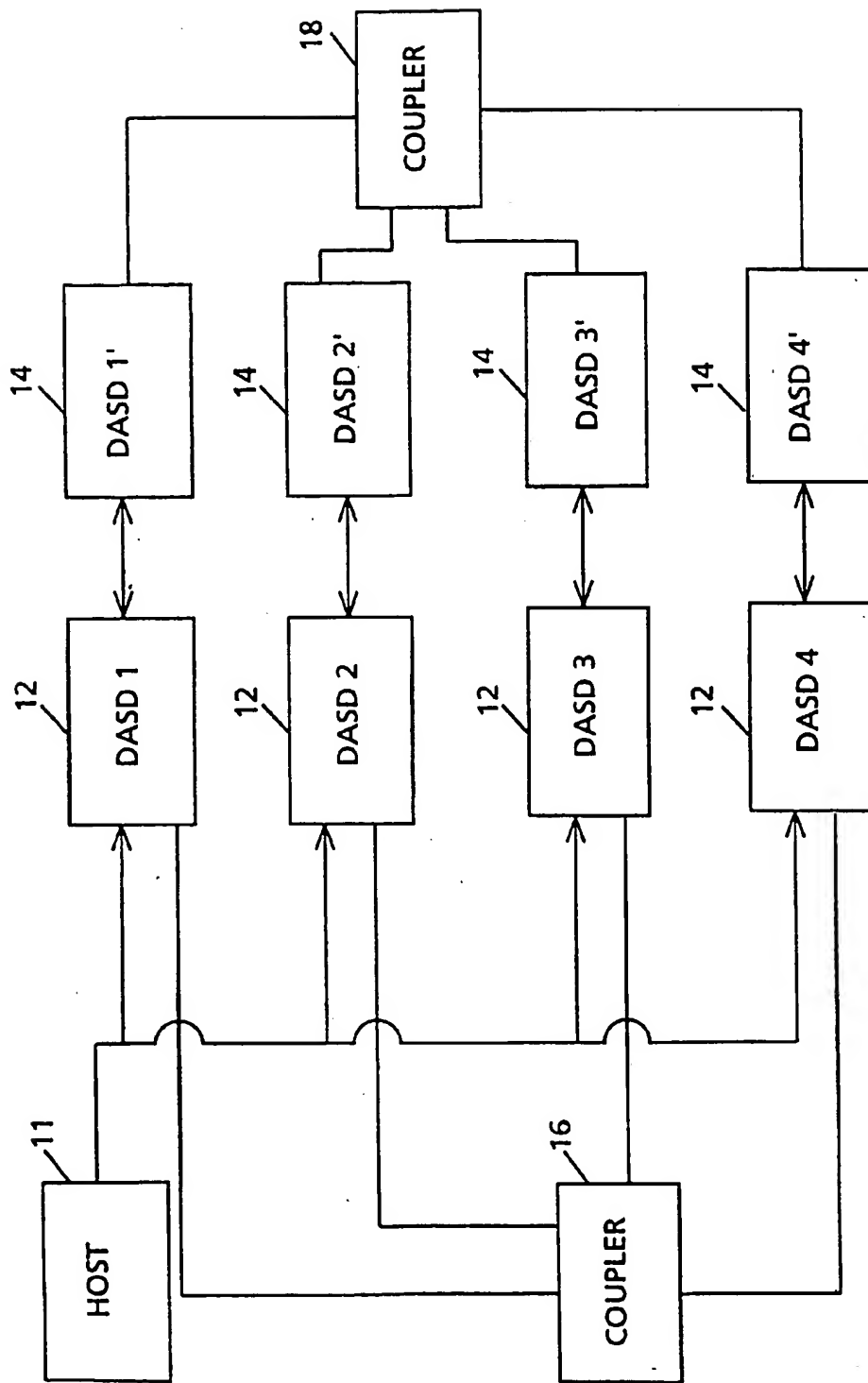
système de la première pluralité de sous-systèmes comprend en outre un moyen pour répartir un signal de séquence sur le site principal, ledit signal de point de contrôle comprenant l'information indiquant la valeur du signal de séquence.

10. Système selon la revendication 8 ou 9, dans lequel le moyen pour envoyer les mises à jour de données et le signal de point de contrôle à chacun des systèmes pendants de la deuxième pluralité de sous-systèmes couplés comprend un moyen qui répond au moyen qui envoie le signal de point de contrôle à chaque système de la première pluralité de sous-systèmes en insérant le signal de point de contrôle dans la séquence de mise à jour de données envoyée par chaque sous-système de ladite première pluralité de sous-systèmes à son pendant.

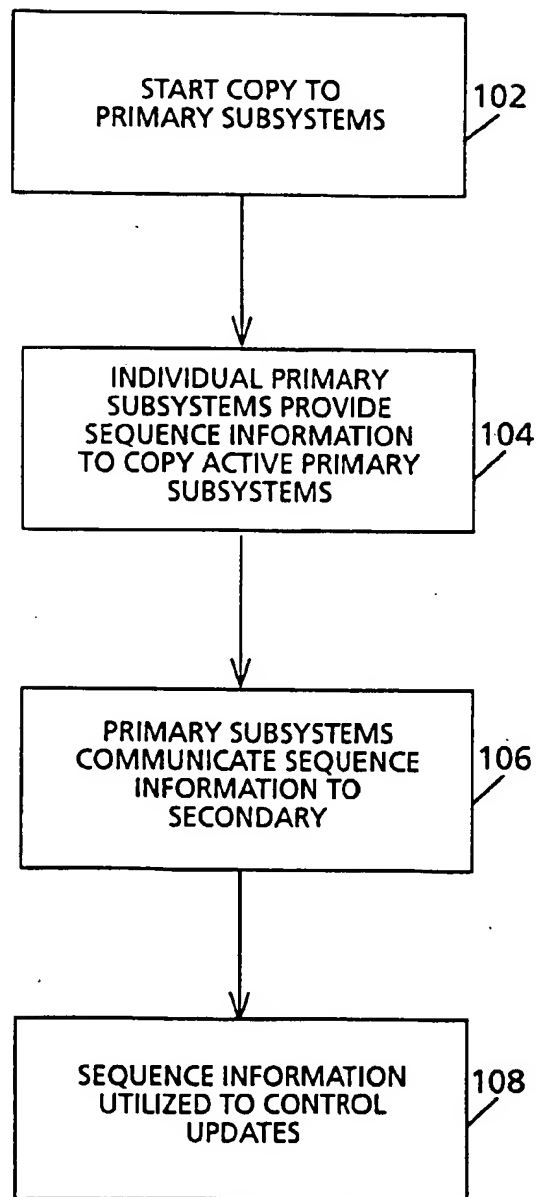


10
PRIOR ART

Figure 1



20
Figure 2



100

Figure 3

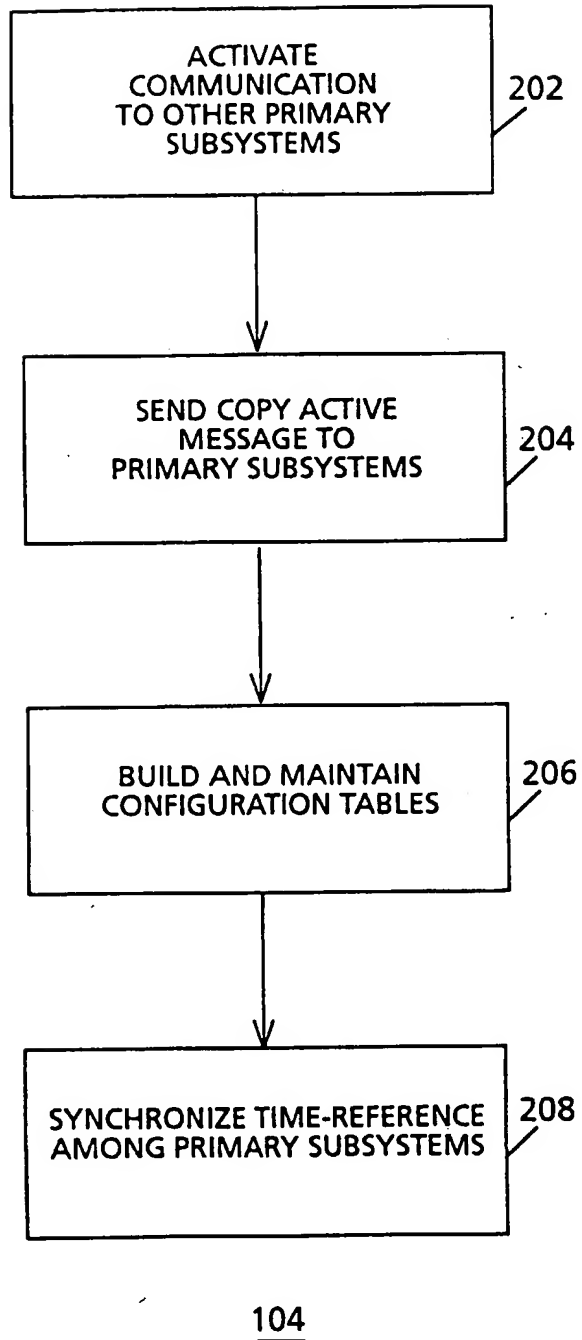


Figure 3A

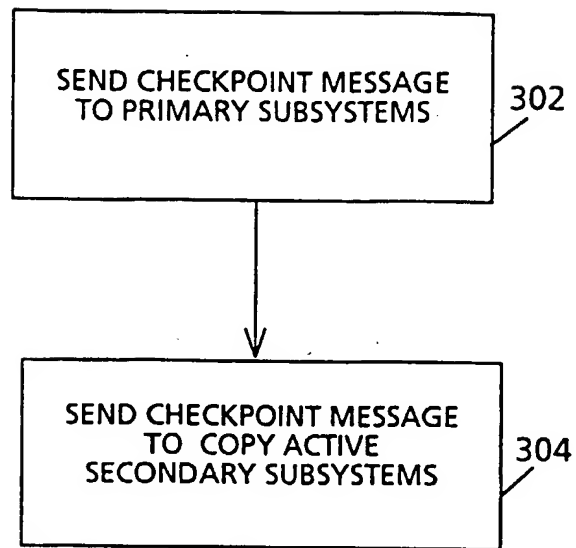
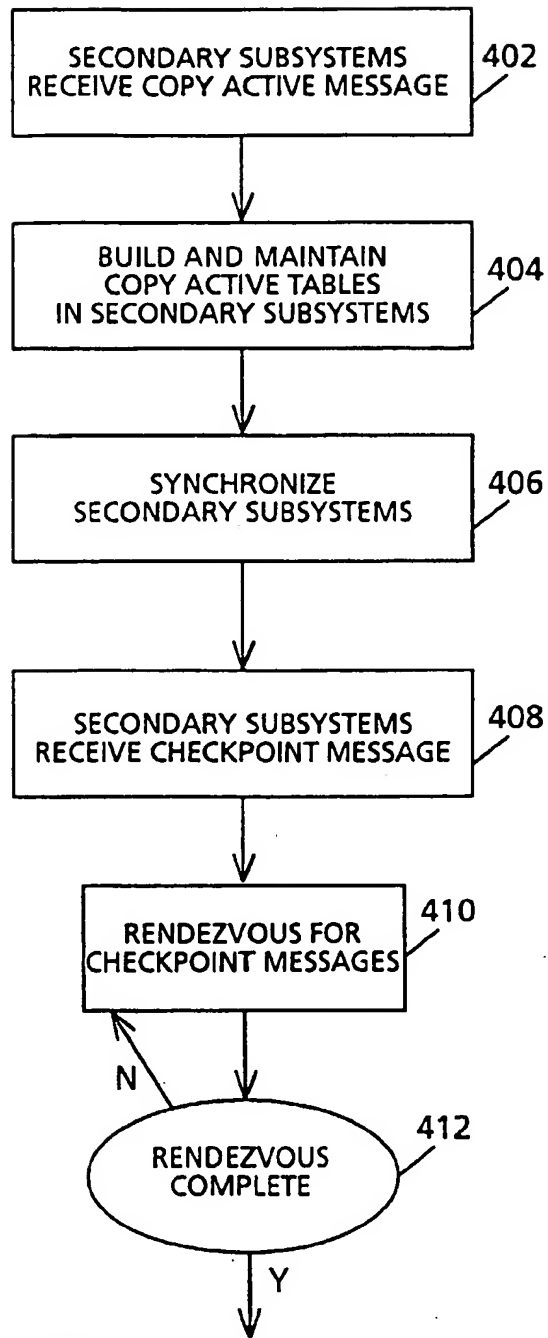
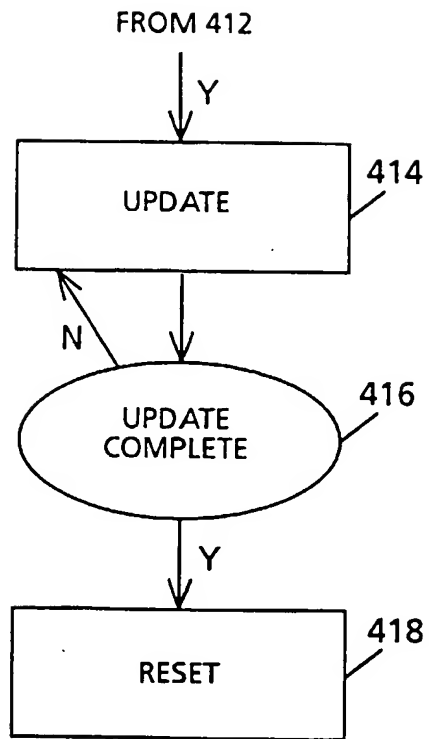


Figure 3B



108

Figure 3C



108

Figure 3D